Voice Recognition: An Enabling Technology For Modern Health Care?

Bryan P. Bergeron, MD
Department of Anesthesia
Massachusetts General Hospital
Boston, MA

Recent performance breakthroughs in affordable, large vocabulary, speaker independent voice recognition systems have rekindled widespread interest in using voice recognition technology to enhance the palatability and effectiveness of clinician-mediated computing. However, even if industry fully addresses the formidable hardware requirements, less than perfect recognition accuracies, discrete voice recognition requirements, and throughput limitations, there are significant cognitive and implementation issues that must be adequately resolved before voice can become a ubiquitous input modality. Cognitive issues include making allowances for individual differences in verbal communication style and skill levels, the relative cognitive load of using a voice enabled interface compared to alternative modalities, and the user's cognitive style. Implementation issues include a significant training requirement, limited portability, lengthy user switching time, questionable privacy, satisfying hardware requirements and the suitability of voice recognition in specific work environments. The inevitable resolution of these issues, coupled with continuously improving voice recognition performance, promises a new era for voice recognition in medicine.

INTRODUCTION

The introduction of affordable, functional voice recognition (VR) systems in the late 1980's spurred a flurry of interest in the use of voice recognition for medical applications[1]. Despite the qualified success of VR in certain niche areas, ranging from aids to the disabled and handicapped to clinical reporting and patient-directed bed positioning[2], the technology behind these early systems was clearly immature[3]. Insufficient recognition accuracy, limited vocabulary size, and overall poor performance, coupled with an inattention to human factors issues, such as an extremely steep learning curve, thwarted the acceptance of voice-enabled applications by the medical community[4-5].

Although commercial voice-enabled medical applications, including clinical reporting systems, have been marketed for several years, they have had only limited success. The most accepted systems tend to be those specifically developed for medical specialties, such as Pathology, where the ability to perform hands-free date entry is a major benefit.

Acceptance of VR seems to require that clinicians first understand the limitations of the technology. For example, while voice-enabled systems have been shown to produce better documentation of patient care[6], they are not only difficult to master but the overall throughput is less than writing by hand[7]. Unfortunately, many vendors have tended to emphasize the former and downplay the latter, in the interest of sales. With few exceptions, the resulting false expectations have generally resulted in clinician frustration and subsequent wholesale rejection of the technology.

While the practical application of VR technology in medicine has been relatively dormant for the past few years, the major VR developers have continued to refine the underlying technology. In addition to creating more powerful VR engines, developers have used human factors studies to help identify interface issues that affect user acceptance of voice-enabled systems. A popular approach in the analysis of VR functionality and performance is the Wizard-of-Oz experimental paradigm, wherein a hidden typist keys in user utterances, mimicking a VR system with a predefined recognition accuracy[8-10]. Using this and other approaches, VR has been tested with variable success to such medically diverse areas as decision support[8,11] and automated history taking from patients[12]. These and similar studies suggest that, while a minimal level of VR performance is critical, it is the user interface in aggregate that defines ease of use and ease of learning, which in turn profoundly affect clinician acceptance.

COGNITIVE ISSUES

VR is better suited for some types of tasks and situations than for others, in part because the technology makes variable demands on how much users must adapt in order to use a voice-enabled system effectively. A difficulty in adapting to voice-enabled systems may be due to individual differences, such as familiarity with computer technology and overall cognitive style, e.g., just as some users are more skilled at navigating through a graphical interface than others, there are marked individual differences in verbal communications skills as well. Similarly, in composing text, some users are linear thinkers, while others tend to be nonlinear, bouncing from topic to topic. These and other individual differences may affect how users interact with and perceive a VR interface versus a typical graphical user interface, given that voice is temporally oriented whereas graphical user interfaces are spatially oriented[13].

Standardized, voice-enabled interface guidelines and the use of temporal metaphors have been suggested as a means of addressing differences in cognitive style[13-15]. For example, tabs can be used to provide different place holders for voiced text, which allows users to think and voice dictate non-linearly and rearrange thoughts later. This and other interface approaches, when properly executed, can support the cognitive styles of a wide range of users. For example, in an interface fully supportive of both the mouse and voice typing, modalities that are inherently supportive of nonlinear and linear thinking, respectively, the degree to which one paradigm is used over the other is entirely user defined. In exchange for this flexibility, the user must learn both the graphical and VR-specific components of an interface.

In addition to issues related to ease of learning, there are other factors that may affect the overall usability of voice-enabled applications in clinical settings. For example, studies relating the relative cognitive load of interacting with a computer by voice while working through significant mental tasks have shown that VR is more effective for some tasks than others[16]. Consider, for example, that it is possible to carry on a conversation with someone while typing. A practiced typist, while generally unaware individual keystrokes, relies on the constant tactile feedback from the keyboard for feedback on accuracy, allowing him or her to look at the monitor only infrequently. In contrast, it is not generally possible to carry on a conversation while working with voice-enabled applications. More importantly, working with voice-enabled systems generally demands that users constantly focus on the display because of their need to verify the accuracy of the voice-to-text translation.

IMPLEMENTATION ISSUES

Key voice-enabled system implementation issues include a significant training requirement, lack of interface portability, a sometimes extended user switching time, demanding hardware requirements, and confidentiality concerns. Arguably the greatest implementation hurdle in introducing a VR-based application into a medical practice is the formidable training requirement. Not only must the VR system be trained to recognize the user's voice, but users must learn to adapt to the VR system requirements. Clinicians, who typically have little time for training, must learn not only the underlying application, but the nuances of the VR interface as well. The investment in training time is often difficult to rationalize, from a personal or departmental perspective, especially when the potential users are either residents rotating through a department or faculty who routinely practice at several hospitals.

Unlike the ubiquitous keyboard, VR systems are usually highly customized and user specific, in that the voice profiles created during system training are generally application, user and hardware specific. For example, simply changing microphones can drastically decrease recognition accuracy. For maximum recognition accuracy, users should restrict themselves to one machine when performing VR work. However, such a limitation may be impractical in a typical hospital setting.

The user-specific files created during the training of a VR system tend to be relatively large, and may take up to a minute or more to load into memory. Because these user-specific files must be loaded before a voice-enabled application can be used, the host machine is effectively unusable during the loading process. When there are multiple users assigned to a single machine in a busy environment, such as an ER, the delay associated with loading user profile may be impractical. Providing individual workstations for each clinician may be economically unfeasible as well. Upgrading to a higher performance computer system, while costly, may prove to be the best solution.

The subtle tapping of a keyboard is now a standard component of any office environment. Keying is both unobtrusive to those working nearby and inherently more secure than speaking to a voice-enabled application. If a telephone can't be used to convey certain information, then a voice-enabled application will be inappropriate as well. Voice-enabled data entry is generally unsuitable when patients are present (e.g., at the bedside).

The hardware requirements of current voice-enabled systems remain a significant impediment to wide-scale use, especially in conservative, fund-limited hospital settings. VR applications are very hardware intensive, in that they demand fast processor speeds, relatively large amounts of dedicated RAM (e.g., over and above system requirements), and larger monitor sizes. Few hospitals can afford to provide all of their staff with 486 or Pentium-class machines with 32 MB or more of RAM and 17" or larger monitors. Interestingly, voice enabled applications require more monitor real estate than graphics-only applications because of the additional space required for the graphical voice controls.

CURRENT STATUS

Given our current understanding of the cognitive and implementation issues associated with VR, what impediments remain that prevent VR from becoming as ubiquitous as the keyboard in medical computing? The solution seems to be related to four key areas: hardware availability and cost; VR software pricing; the design of standardized, VR-aware user interfaces; and VR system throughput. The first impediment, hardware availability and cost, will be solved with time. The continued, precipitous drop in hardware prices, hopefully accompanied by decreases in RAM prices, will continue to lower the entry cost for small clinics and individual clinicians. However, the departmental and hospital-wide application of VR technology may take several years, given the relative abundance of proprietary operating systems in these settings. The majority of VR vendors support only DOS or Windows, and it is unlikely that hospitals will migrate their existing systems to these operating systems simply to comply with VR requirements.

Compared to most mature, mass-marketed computer technologies, VR technology is still a premium commodity. General-purpose VR engines are available for about 1/4 the price of price of a typical desktop computer, while voice enabled medical applications are often sold for significantly more than the basic application price. Given the marketing attention focused on VR technology, prices of both general-purpose and medical-specific VR systems should drop precipitously in the upcoming year or two.

User interface specifications are beginning to reflect the potential role of VR as a major component of multimodal interfaces. Microsoft, IBM, and other major vendors now provide support for VR through hardware or operating system level calls. The move to a true multimodal interface design is not universal, however. For example, Web browsers are notoriously difficult to voice enable, having been designed primarily for mouse-directed point-and-click interaction.

Perhaps the greatest hurdle, and one that represents the complex interaction of hardware performance, user interface design optimization, as well as the performance of the basic VR engine, is that of effective system throughput. Whereas the maximum theoretical number of words that can be recognized per minute is limited by the responsiveness of the hardware and the accuracy of the VR engine, the actual throughput is further limited by how well text manipulation operations are supported in particular voice-enabled applications. For example, in free text voice-enabled dictation applications, actual throughput is greatly dependent on the demands placed on the user for locating and correcting errors, which is in turn related to factors such as hardware performance and user interface design.

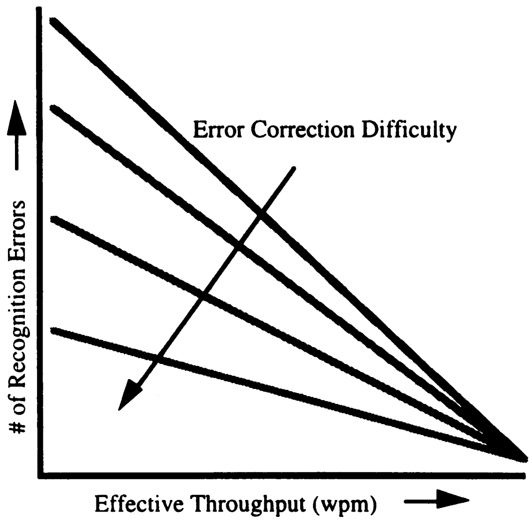

Figure 1. Family of Curves for Effective Throughput vs # of Recognition Errors

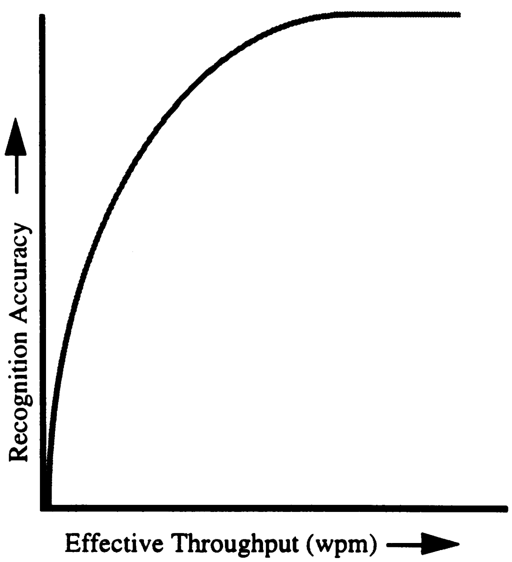

Figure 2. Recognition Accuracy vs Effective Throughput

Figure 1 illustrates the general relationship between effective throughput, the number of recognition errors, and the difficulty in locating and correcting errors. As the number of errors increases, actual throughput declines. In addition, the depression in throughput is more pronounced as the number of errors increases[17]. This effect may be due to cognitive factors, such as elevated user frustration at higher error levels.

Cognitive factors likely affect the relationship between effective throughput and recognition accuracy. For example, as illustrated in Figure 2, there is a non-linear relationship between effective throughput and recognition accuracy[17]. With higher recognition accuracies, small improvements in recognition accuracy result in marked improvements in effective throughput. Conversely, at lower recognition accuracies, a relatively small increase in recognition accuracy is associated with only a small increase in effective throughput. That is, a change from three to two recognition errors per hundred words is less noticeable than a change from 16 to 15 errors per hundred words.

DISCUSSION

The failures experienced with early voice-enabled medical applications were certainly due in part to the immaturity of the VR technology, and in part to a general lack of understanding among developers of how and when to best use voice as an input modality. Subsequent improvements in VR technology, and computer technology in general, have removed major barriers to the widespread adoption of voice-enabled applications in medicine. What remains to be developed are more effective voice-aware user and application interfaces, as well as a better understanding of when to use VR to solve specific problems. Although VR has had the stigma of an emergent technology in search of an application, the current prognosis seems exceedingly bright.

References

1. Bergeron BP, Locke S. Speech recognition as a user interface. *MD Comput* 1990;7(5):329-34.
2. Liang MD, Narayanan K. The application of voice recognition to robotic positioning of a hospital bed. *Speech Technol* 1989;5(1):30-3.

3. Forren M, Mitchell C. Voice input in real time decision making. *Proceedings of the 1986 IEEE International Conference on Systems, Man, and Cybernetics*: IEEE, 1986:879-84.
4. Murchie C, Kenny G. Comparison of keyboard, light pen and voice recognition as methods of data input. *Intl J of Clin Monitoring and Computing* 1988;5(4):243-6.
5. Jones D, Frankish C, Taylor M, Starr A, Richardson I. Matching the machine to man: the human factors issues in voice input. voice processing: technology and opportunity in the office. London, UK: *Online Publications*, 1985:23-30.
6. Bunschoten B. What role will speech recognition play in health care? *Health Data Management* 1996;4(1):38-41.
7. Linn N, Rubenstein R, Bowler A, Dixon J. Improving the quality of emergency department documentation using the voice-activated word processor: interim results. *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care*, 1993:772-6.
8. Detmer W, Shiffman S, Wyatt J, Friedman C, Lane C, Fagan L. A continuous-speech interface to a decision support system: An evaluation using a Wizard-of-Oz experimental paradigm. *JAMIA* 1995;2(1):46-57.
9. Oviatt S. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 1995;9(1):19-35.
10. Love S, Foster J, Jack M. On the effects of individual differences with reference to spoken language dialogue systems. Vienna Conference, *VCHCI'93*. Vienna, Austria: Springer-Verlag, 1993:405-6.
11. Shiffman S, Lane C, Johnson K, Fagan L. The integration of a continuous-speech-recognition system with the QMR diagnostic program. *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*, 1992:767-71.
12. Johnson K, Poon A, Shiffman S, Lin R, Fagan L. A history-taking system that uses continuous speech recognition. *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*, 1992:757-61.
13. Johnson R. Labs explore aural user interfaces. *Electronic Engineering Times* 1993(763):33-6.
14. Gardner-Bonneau D. Human factors problems in interactive voice response allocations: do we need a guideline/standard? *Proceedings of the human factors society 36th annual meeting: innovations for interaction*. Atlanta, GA: Human Factors Society, 1992:222-6.
15. Gosbee J, Clay M. Human factors problem analysis of a voice-recognition computer-based medical record. *Proceedings of the Sixth Annual IEEE Symposium on Computer-Based Medical Systems*. Ann Arbor, MI: IEEE Computer Society Press, 1993:235-40.
16. Murata A. Experimental discussion on effectiveness of voice input in a dual task situation. *Transactions of the Institute of Electronics, Information and Communication Engineers* 1995;78(6):982-8.
17. Bergeron B. Unpublished study.